Probing the space of grammatical variation: induction of cross-lingual grammatical constraints from treebanks

Felice Dell'Orletta
Università di Pisa, Dipartimento di
Informatica - Largo B. Pontecorvo 3
ILC-CNR - via G. Moruzzi 1
56100 Pisa, Italy
felice.dellorletta@ilc.cnr.it

Alessandro Lenci
Università di Pisa, Dipartimento di
Linguistica - Via Santa Maria 36
56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Simonetta Montemagni
ILC-CNR - via G. Moruzzi 1
56100 Pisa, Italy
simonetta.montemagni@ilc.cnr.it

Vito Pirrelli
ILC-CNR - via G. Moruzzi 1
56100 Pisa, Italy
vito.pirrelli@ilc.cnr.it

Abstract

The paper reports on a detailed quantitative analysis of distributional language data of both Italian and Czech, highlighting the relative contribution of a number of distributed grammatical factors to sentence-based identification of subjects and direct objects. The work uses a Maximum Entropy model of stochastic resolution of conflicting grammatical constraints and is demonstrably capable of putting explanatory theoretical accounts to the test of usage-based empirical verification.

1 Introduction

The paper illustrates the application of a *Maximum Entropy* (henceforth MaxEnt) model (Ratnaparkhi 1998) to the processing of subjects and direct objects in Italian and Czech. The model makes use of richly annotated Treebanks to determine the types of linguistic factors involved in the task and weigh up their relative salience. In doing so, we set ourselves a two-fold goal. On the one hand, we intend to discuss the use of Treebanks to discover typologically relevant and linguistically motivated factors and assess the relative contribution of the latter to cross-linguistic parsing issues. On the other hand, we are interested in testing the empirical plausibility of constraint-resolution models of language processing (see infra) when confronted with real language data.

Current research in natural language learning and processing supports the view that grammatical competence consists in mastering and integrating multiple, parallel constraints (Seidenberg and MacDonald 1999, MacWhinney 2004). Moreover, there is growing consensus on two major properties of grammatical constraints: i.) they are probabilistic "soft constraints" (Bresnan *et al.* 2001), and ii.) they have an inherently functional nature, involving different types of linguistic (and non linguistic) information (syntactic, semantic, etc.). These features emerge particularly clearly in dealing with one of the core aspects of grammar learning: the ability to identify *syntactic relations* in text. Psycholinguistic evidence shows that speakers learn to identify sentence subjects and direct objects by combining various types of probabilistic, functional cues, such as word order, noun animacy, definiteness, agreement, etc. An important observation is that the relative prominence of each such cue can considerably vary cross-linguistically. Bates *et al.* (1984), for example, argue that while, in English, word order is the most effective cue for Subject-Object Identification (henceforth *SOI*) both in syntactic processing and during the child's syntactic development, the same cue plays second fiddle in relatively free phrase-order languages such as Italian or German.

If grammatical constraints are inherently probabilistic (Manning 2003), the path through which adult grammar competence is acquired can be viewed as the process of building a stochastic model out of the linguistic input. In computational linguistics, MaxEnt models have

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 21–28,
Sydney, July 2006. ©2006 Association for Computational Linguistics

proven to be robust statistical learning algorithms that perform well in a number of processing tasks. Being supervised learning models, they require richly annotated data as training input. Before we turn to the use of Treebanks for training a MaxEnt model for *SOI*, we first analyse the range of linguistic factors that are taken to play a significant role in the task.

2 Subjects and objects in Czech and Italian

Grammatical relations - such as subject (*S*) and direct object (*O*) - are variously encoded in languages, the two most widespread strategies being: i) structural encoding through *word order*, and ii) morpho-syntactic marking. In turn, morpho-syntactic marking can apply either on the noun head only, in the form of *case inflections*, or on both the noun and the verb, in the form of agreement marking (Croft 2003). Besides formal coding, the distribution of subjects and object is also governed by semantic and pragmatic factors, such as noun animacy, definiteness, topicality, etc. As a result, there exists a variety of linguistic clues jointly co-operating in making a particular noun phrase the subject or direct object of a sentence. Crucially for our present purposes, cross-linguistic variation does not only concern the particular strategy used to encode *S* and *O*, but also the *relative strength* that each factor plays in a given language. For instance, while English word order is by and large the dominant clue to identify *S* and *O*, in other languages the presence of a rich morphological system allows word order to have a much looser connection with the coding of grammatical relations, thus playing a secondary role in their identification. Moreover, there are languages where semantic and pragmatic constraints such as animacy and/or definiteness play a predominant role in the processing of grammatical relations. A large spectrum of variations exists, ranging from languages where *S must* have a higher degree of animacy and/or definiteness relative to *O*, to languages where this constraint only takes the form of a softer statistical preference (cf. Bresnan *et al.* 2001).

The goal of this paper is to explore the area of this complex space of grammar variation through careful assessment of the distribution of *S* and *O* tokens in Italian and Czech. For our present analysis, we have used a MaxEnt statistical model trained on data extracted from two syntactically annotated corpora: the *Prague Dependency Treebank* (PDT, Bohmova *et al.* 2003) for Czech, and the *Italian Syntactic Semantic Treebank* (ISST, Montemagni *et al.* 2003) for Italian. These corpora have been chosen not only because they are the largest syntactically annotated resources for the two languages, but also because of their high degree of comparability, since they both adopt a dependency-based annotation scheme.

Czech and Italian provide an interesting vantage point for the cross-lingual analysis of grammatical variation. They are both Indo-European languages, but they do not belong to the same family: Czech is a West Slavonic language, while Italian is a Romance language. For our present concerns, they appear to share two crucial features: i) the free order of grammatical relations with respect to the verb; ii) the possible absence of an overt subject. Nevertheless, they also greatly differ due to: the virtual non-existence of case marking in Italian (with the only marginal exception of personal pronouns), and the degree of phrase-order freedom in the two languages. Empirical evidence supporting the latter claim is provided in Table 1, which reports data extracted from PDT and ISST. Notice that although in both languages *S* and *O* can occur either pre-verbally or post-verbally, Czech and Italian greatly differ in their propensity to depart from the (unmarked) SVO order. While in Italian preverbal *O* is highly infrequent (1.90%), in Czech more than 30% of *O* tokens occur before the verb. The situation is similar but somewhat more balanced in the case of *S*, which occurs post-verbally in 22.21% of the Italian cases, and in 40% of Czech ones. For sure, one can argue that, in spoken Italian, the number of pre-verbal objects is actually higher, because of the greater number of left dislocations and topicalizations occurring in informal speech. However reasonable, the observation does not explain away the distributional differences in the two corpora, since both PDT and ISST contain written language only. We thus suggest that there is clear empirical evidence in favour of a systematic, higher phrase-order freedom in Czech, arguably related to the well-known correlation of Czech constituent placement with sentence information structure, with the element carrying new information showing a tendency to occur sentence-finally (Stone 1990). For our present concerns, however, aspects of information structure, albeit central in Czech grammar, were not taken into account, as they happen not to be

		Czech		Italian	
		Subj	Obj	Subj	Obj
Pos	Pre	59.82%	30.27%	77.79%	1.90%
	Post	40.18%	69.73%	22.21%	98.10%
	All	100.00%	100.00%	100.00%	100.00%
Agr	Agr	98.50%	56.54%	97.73%	58.33%
	NoAgr	1.50%	43.46%	2.27%	41.67%
	All	100.00%	100.00%	100.00%	100.00%
Anim	Anim	34.10%	15.42%	50.18%	10.67%
	NoAnim	65.90%	84.58%	49.82%	89.33%
	All	100.00%	100.00%	100.00%	100.00%

Table 1 –*Distribution of Czech and Italian S and O wrt word order, agreement and noun animacy*

	Czech	
	Subj	Obj
Nominative	53.83%	0.65%
Accusative	0.15%	28.30%
Dative	0.16%	9.54%
Genitive	0.22%	2.03%
Instrumental	0.01%	3.40%
Ambiguous	45.63%	56.08%
All	100.00%	100.00%

Table 2 - *Distribution of Czech S and O wrt case*

marked-up in the Italian corpus.

According to the data reported in Table 1, Czech and Italian show similar correlation patterns between animacy and grammatical relations. *S* and *O* in ISST were automatically annotated for animacy using the SIMPLE Italian computational lexicon (Lenci *et al.* 2000) as a background semantic resource. The annotation was then checked manually. Czech *S* and *O* were annotated for animacy using Czech WordNet (Pala and Smrz 2004); it is worth remarking that in Czech animacy annotation was done only automatically, without any manual revision. Italian shows a prominent asymmetry in the distribution of animate nouns in subject and object roles: over 50% of ISST subjects are animate, while only 10% of the objects are animate. Such a trend is also confirmed in Czech – although to a lesser extent - with 34.10% of animate subjects vs. 15.42% of objects.[1] Such an overwhelming preference for animate subjects in corpus data suggests that animacy may play a very important role for *S* and *O* identification in both languages.

Corpus data also provide interesting evidence concerning the actual role of morpho-syntactic constraints in the distribution of grammatical relations. *Prima facie*, agreement and case are the strongest and most directly accessible clues for *S/O* processing, as they are marked both overtly and locally. This is also confirmed by psycholinguistic evidence, showing that subjects tend to rely on these clues to identify *S/O*. However, it should be observed that agreement can be relied upon conclusively in *S/O* processing only when a nominal constituent and a verb do not agree in number and/or person (as in *leggono il libro* '(they) read the book'). Conversely, when N and V share the same person and number, no conclusion can be drawn, as trivially shown by a sentence like *il bambino legge il libro* 'the child reads the book'. In ISST, more than 58% of *O* tokens agree with their governing V, thus being formally indistinguishable from *S* on the basis of agreement features. PDT also exhibits a similar ratio, with 56% of *O* tokens agreeing with their verb head. Analogous considerations apply to case marking, whose perceptual reliability is undermined by morphological syncretism, whereby different cases are realized through the same marker. Czech data reveal the massive extent of this phenomenon and its impact on *SOI*. As reported in Table 2, more than 56% of *O* tokens extracted from PDT are formally indistinguishable from *S* in case ending. Similarly, 45% of *S* tokens are formally indistinguishable from *O* uses on the same ground. All in all, this means that in 50% of the cases a Czech noun can not be understood as the *S/O* of a sentence by relying on overt case marking only.

To sum up, corpus data lend support to the idea that in both Italian and in Czech *SOI* is governed by a complex interplay of probabilistic constraints of a different nature (morpho-syntactic, semantic, word order, etc.) as the latter are neither singly necessary nor jointly sufficient to attack the processing task at hand. It is tempting to hypothesize that the joint distribution of these data can provide a statistically reliable basis upon which relevant probabilistic constraints are bootstrapped and combined consistently. This should be possible due to i) the different degrees of clue salience in the two languages and ii) the functional need to minimize

[1] In fact, the considerable difference in animacy distribution between the two languages might only be an artefact of the way we annotated Czech nouns semantically, on the basis of their context-free classification in the Czech WordNet.

processing ambiguity in ordinary communicative exchanges. With reference to the latter point, for example, we may surmise that a speaker will be more inclined to violate one constraint on *S/O* distribution (e.g. word order) when another clue is available (e.g. animacy) that strongly supports the intended interpretation only. The following section illustrates how a MaxEnt model can be used to model these intuitions by bootstrapping constraints and their interaction from language data.

3 Maximum Entropy modelling

The MaxEnt framework offers a mathematically sound way to build a probabilistic model for *SOI*, which combines different linguistic cues. Given a linguistic context *c* and an outcome $a \in A$ that depends on *c*, in the MaxEnt framework the conditional probability distribution $p(a|c)$ is estimated on the basis of the assumption that no *a priori* constraints must be met other than those related to a set of features $f_j(a,c)$ of *c*, whose distribution is derived from the training data. It can be proven that the probability distribution *p* satisfying the above assumption is the one with the highest entropy, is unique and has the following exponential form (Berger *et al.* 1996):

$$(1) \qquad p(a\,|\,c) = \frac{1}{Z(c)} \prod_{j=1}^{k} a_j^{f_j(a,c)}$$

where $Z(c)$ is a normalization factor, $f_j(a,c)$ are the values of *k* features of the pair (a,c) and correspond to the linguistic cues of *c* that are relevant to predict the outcome *a*. Features are extracted from the training data and define the constraints that the probabilistic model *p* must satisfy. The parameters of the distribution $\alpha_1, \ldots, \alpha_k$ correspond to *weights* associated with the features, and determine the relevance of each feature in the overall model. In the experiments reported below feature weights have been estimated with the Generative Iterative Scaling (GIS) algorithm implemented in the AMIS software (Miyao and Tsujii 2002).

We model *SOI* as the task of predicting the correct syntactic function $\varphi \in$ {*subject*, *object*} of a noun occurring in a given syntactic context σ. This is equivalent to building the conditional probability distribution $p(\varphi|\sigma)$ of having a syntactic function φ in a syntactic context σ. Adopting the MaxEnt approach, the distribution *p* can be rewritten in the parametric form of (1), with features corresponding to the linguistic contextual cues relevant to *SOI*. The context σ is a pair $<v_\sigma, n_\sigma>$, where v_σ is the verbal head and n_σ its nominal dependent in σ. This notion of σ departs from more traditional ways of describing an *SOI* context as a triple of one verb and two nouns in a certain syntactic configuration (e.g, *SOV* or *VOS*, etc.). In fact, we assume that *SOI* can be stated in terms of the more local task of establishing the grammatical function of a noun *n* observed in a verb-noun pair. This simplifying assumption is consistent with the claim in MacWhinney *et al.* (1984) that *SVO* word order is actually derivative from *SV* and *VO* local patterns and downplays the role of the transitive complex construction in sentence processing. Evidence in favour of this hypothesis also comes from corpus data: for instance, in ISST complete subject-verb-object configurations represent only 26% of the cases, a small percentage if compared to the 74% of verb tokens appearing with either a subject or an object only; a similar situation can be observed in PDT where complete subject-verb-object configurations occur in only 20% of the cases. Due to the comparative sparseness of canonical *SVO* constructions in Czech and Italian, it seems more reasonable to assume that children should pay a great deal of attention to both *SV* and *VO* units as cues in sentence perception (Matthews *et al.* in press). Reconstruction of the whole lexical *SVO* pattern can accordingly be seen as the end point of an acquisition process whereby smaller units are re-analyzed as being part of more comprehensive constructions. This hypothesis is more in line with a *distributed* view of canonical constructions as derivative of more basic local positional patterns, working together to yield more complex and abstract constructions. Last but not least, assuming verb-noun pairs as the relevant context for *SOI* allows us to simultaneously model the interaction of word order variation with pro-drop.

4 Feature selection

The most important part of any MaxEnt model is the selection of the context features whose weights are to be estimated from data distributions. Our feature selection strategy is grounded on the main assumption that features should correspond to theoretically and typologically well-motivated contextual cues. This allows us to evaluate the probabilistic model also with respect to its consistency with current linguistic generalizations. In turn, the model can be used as a probe into the correspondence between theoretically motivated generalizations and usage-based empirical evidence.

Features are binary functions $f_{k_i,\varphi}$ (φ,σ), which test whether a certain cue k_i for the feature φ occurs in the context σ. For our MaxEnt model, we have selected different features types that test morpho-syntactic, syntactic, and semantic key dimensions in determining the distribution of *S* and *O*.

Morpho-syntactic features. These include N-V agreement, for Italian and Czech, and case, only for Czech. The combined use of such features allow us not only to test the impact of morpho-syntactic information on *SOI*, but also to analyze patterns of cross-lingual variation stemming from language specific morphological differences, e.g. lack of case marking in Italian.

Word order. This feature essentially test the position of the noun wrt the verb, for instance:

$$(2)\ f_{post,subj}(subj, s) = \begin{cases} 1 & if\ noun_s.pos = post \\ 0 & otherwise \end{cases}$$

Animacy. This is the main semantic feature, which tests whether the noun in σ is animate or inanimate (cf. section 2). The centrality of this cue for grammatical relation assignment is widely supported by typological evidence (cf. Aissen 2003, Croft 2003). The Animacy Markedness Hierarchy - representing the relative markedness of the associations between grammatical functions and animacy degrees – is actually assigned the role of a functional universal principle in grammar. The hierarchy is reported below, with each item in these scales been less marked than the elements to its right:

Animacy Markedness Hierarchy
Subj/Human > Subj/Animate > Subj/Inanimate
Obj/Inanimate > Obj/Animate > Obj/Human

Markedness hierarchies have also been interpreted as probabilistic constraints estimated from corpus data (Bresnan *et al.* 2001). In our MaxEnt model we have used a reduced version of the animacy markedness hierarchy in which human and animate nouns have been both subsumed under the general class animate.

Definiteness tests the degree of “referentiality” of the noun in a context pair σ. Like for animacy, definiteness has been claimed to be associated with grammatical functions, giving rise to the following universal markedness hierarchy Aissen (2003):

Definiteness Markedness Hierarchy
Subj/Pro > Subj/Name > Subj/Def > Subj/Indef
Obj/Indef > Obj/Def > Obj/Name > Obj/Pro

According to this hierarchy, subjects with a low degree of definiteness are more marked than subjects with a high degree of definiteness (for objects the reverse pattern holds). Given the importance assigned to the definiteness markedness hierarchy in current linguistic research, we have included the definiteness cue in the MaxEnt model. In our experiments, for Italian we have used a compact version of the definiteness scale: the definiteness cue tests whether the noun in the context pair i) is a name or a pronoun ii) has a definite article iii), has an indefinite article or iv) is a bare noun (i.e. with no article). It is worth saying that bare nouns are usually placed at the bottom end of the definiteness scale. Since in Czech there is no article, we only make a distinction between proper names and common nouns.

5 Testing the model

The Italian MaxEnt model was trained on 14,643 verb-subject/object pairs extracted from ISST. For Czech, we used a training corpus of 37,947 verb-subject/object pairs extracted from PDT. In both cases, the training set was obtained by extracting all verb-subject and verb-object dependencies headed by an active verb, with the exclusion of all cases where the position of the nominal constituent was grammatically determined (e.g. clitic objects, relative clauses). It is interesting to note that in both training sets the proportion of subjects and objects relations is nearly the same: 63.06%-65.93% verb-subject pairs and 36.94%-34.07% verb-object pairs for Italian and Czech respectively.

The test corpus consists of a set of verb-noun pairs randomly extracted from the reference Treebanks: 1,000 pairs for Italian and 1,373 for Czech. For Italian, 559 pairs contained a subject and 441 contained an object; for Czech, 905 pairs contained a subject and 468 an object. Evaluation was carried out by calculating the percentage of correctly assigned relations over the total number of test pairs (accuracy). As our model always assigns one syntactic relation to each test pair, accuracy equals both standard precision and recall.

	Czech		Italian	
	Subj	Obj	Subj	Obj
Preverb	1.99%	19.40%	0.00%	6.90%
Postverb	71.14%	7.46%	71.55%	21.55%
Anim	0.50%	3.98%	6.90%	21.55%
Inanim	72.64%	22.89%	64.66%	6.90%
Nomin	0.00%	1.00%	Na	
Genitive	0.50%	0.00%		
Dative	1.99%	0.00%		
Accus	0.00%	0.00%		
Instrum	0.00%	0.00%		
Ambig	70.65%	25.87%		
Agr	70.15%	25.87%	61.21%	12.07%
NoAgr	2.99%	0.50%	7.76%	1.72%
NAAgr	0.00%	0.50%	2.59%	14.66%

Table 3 – *Types of errors for Czech and Italian*

	Czech		Italian	
	Subj	Obj	Subj	Obj
Preverb	1.24E+00	5.40E-01	1.31E+00	2.11E-02
Postverb	8.77E-01	1.17E+00	5.39E-01	1.38E+00
Anim	1.16E+00	6.63E-01	1.28E+00	3.17E-01
Inanim	1.03E+00	9.63E-01	8.16E-01	1.23E+00
PronName	1.13E+00	7.72E-01	1.13E+00	8.05E-01
DefArt	1.05E+00	9.31E-01	1.01E+00	1.02E+00
IndefArt			6.82E-01	1.26E+00
NoArticle			9.91E-01	1.02E+00
Nomin	1.23E+00	2.22E-02	Na	
Genitive	2.94E-01	1.51E+00		
Dative	2.85E-02	1.49E+00		
Accus	8.06E-03	1.39E+00		
Instrum	3.80E-03	1.39E+00		
Agr	1.18E+00	6.67E-01	1.28E+00	4.67E-01
NoAgr	7.71E-02	1.50E+00	1.52E-01	1.58E+00
NAAgr	3.75E-01	1.53E+00	2.61E-01	1.84E+00

Table 4 - *Feature value weights in NLC for Czech and Italian*

We have assumed a baseline score of 56% for Italian and of 66% for Czech, corresponding to the result yielded by a naive model assigning to each test pair the most frequent relation in the training corpus, i.e. subject. Experiments were carried out with the general features illustrated in section 4: verb agreement, case (for Czech only), word order, noun animacy and noun definiteness.

Accuracy on the test corpus is 88.4% for Italian and 85.4% for Czech. A detailed error analysis for the two languages is reported in Table 3, showing that in both languages subject identification appears to be particularly problematic. In Czech, it appears that the prototypically mistaken subjects are post-verbal (71.14%), inanimate (72.64%), ambiguously case-marked (70.65%) and agreeing with the verb (70.15%), where reported percentages refer to the whole error set. Likewise, Italian mistaken subjects can be described thus: they typically occur in post-verbal position (71.55%), are mostly inanimate (64.66%) and agree with the verb (61.21%). Interestingly, in both languages, the highest number of errors occurs when a) N has the least prototypical syntactic and semantic properties for *O* or *S* (relative to word order and noun animacy) and b) morpho-syntactic features such as agreement and case are neutralised. This shows that MaxEnt is able to home in on the core linguistic properties that govern the distribution of *S* and *O* in Italian and Czech, while remaining uncertain in the face of somewhat peripheral and occasional cases.

A further way to evaluate the goodness of fit of our model is by inspecting the weights associated with feature values for the two languages. They are reported in Table 4, where grey cells highlight the preference of each feature value for either subject or object identification. In both languages agreement with the verb strongly relates to the subject relation. For Czech, nominative case is strongly associated with subjects while the other cases with objects. Moreover, in both languages preverbal subjects are strongly preferred over preverbal objects; animate subjects are preferred over animate objects; pronouns and proper names are typically subjects.

Let us now try to relate these feature values to the Markedness Hierarchies reported in section 4. Interestingly enough, if we rank the Italian *Anim* and *Inanim* values for subjects and objects, we observe that they distribute consistently with the *Animacy Markedness Hierarchy*: *Subj*/*Anim* > *Subj*/*Inanim* and *Obj*/*Inanim* > *Obj*/*Anim*. This is confirmed by the Czech results. Similarly, by ranking the Italian values for the definiteness features in the *Subj* column by decreasing weight values we obtain the following ordering: *PronName* > *DefArt* > *IndefArt* > *NoArt*, which nicely fits in with the *Definiteness Markedness Hierarchy* in section 4. The so-called "markedness reversal" is replicated with a good degree of approximation, if we focus on the values for the same features in the *Obj* column: the *PronName* feature represents the most marked option, followed by *IndefArt*, *DefArt* and *NoArt* (the latter two showing the same feature value). The exception here is represented by the relative ordering of *IndefArt* and *DefArt* which however show very close values. The same seems to hold for Czech, where the feature ordering for *Subj* is *PronName* > *DefArt*/*IndefArt*/*NoArt* and the reverse is observed for *Obj*.

5.1 Evaluating comparative feature salience

The relative salience of the different constraints acting on *SOI* can be inferred by comparing the weights associated with individual feature values. For instance, Goldwater and Johnson (2003) show that MaxEnt can successfully be applied to learn constraint rankings in Optimality Theory, by assuming the parameter weights <α1, …, αk> as the ranking values of the constraints.

Table 5 illustrates the constraint ranking for the two languages, ordered by decreasing weight values for both *S* and *O*. Note that, although not all constraints are applicable in both languages, the weights associated with applicable constraints exhibit the same relative salience in Czech and Italian. This seems to suggest the existence of a rather dominant (if not universal) salience scale of *S* and *O* processing constraints, in spite of the considerable difference in the marking strategies adopted by the two languages. As the relative weight of each constraint crucially depends on its overall interaction with other constraints on a given processing task, absolute weight values can considerably vary from language to language, with a resulting impact on the distribution of *S* and *O* constructions. For example, the possibility of overtly and unambiguously marking a direct object with case inflection makes wider room for preverbal use of objects in Czech. Conversely, lack of case marking in Italian considerably limits the preverbal distribution of direct objects. This evidence, however, appears to be an epiphenomenon of the interaction of fairly stable and invariant preferences, reflecting common functional tendencies in language processing. As shown in Table 5, if constraint ranking largely confirms the interplay between animacy and word order in Italian, Czech does not contradict it but rather re-modulate it somewhat, due to the "perturbation" factors introduced by its richer battery of case markers.

6 Conclusions

Probabilistic language models, machine language learning algorithms and linguistic theorizing all appear to support a view of language processing as a process of dynamic, on-line resolution of conflicting grammatical constraints. We begin to gain considerable insights into the complex process of bootstrapping nature and behaviour of these constraints upon observing their actual distribution in perceptually salient contexts. In our view of things, this trend outlines a promising framework providing fresh support to usage-based models of language acquisition through mathematical and computational simulations. Moreover, it allows scholars to investigate patterns of cross-linguistic typological variation that crucially depend on the appropriate setting of model parameters. Finally, it promises to solve, on a principled basis, traditional performance-oriented *cruces* of grammar theorizing such as degrees of human acceptability of ill-formed grammatical constructions (Hayes 2000) and the inherently graded compositionality of linguistic constructions such as morpheme-based words and word-based phrases (Bybee 2002, Hay and Baayen 2005).

We argue that the current availability of comparable, richly annotated corpora and of mathematical tools and models for corpus exploration make time ripe for probing the space of grammatical variation, both intra- and inter-linguistically, on unprecedented levels of sophistication and granularity. All in all, we anticipate that such a convergence is likely to have a twofold impact: it is bound to shed light on the integration of performance and competence factors in language study; it will make mathematical models of language increasingly able to accommodate richer and richer language evidence, thus putting explanatory theoretical accounts to the test of a usage-based empirical verification.

In the near future, we intend to pursue two parallel lines of development. First we would like to increase the context-sensitiveness of our processing task by integrating binary grammatical constraints into the broader context of multiply conflicting grammar relations. This way, we will be in a position to capture the constraint that a (transitive) verb has at most one subject and one object, thus avoiding multiple assignment of subject (object) relations in the same context. Suppose, for example, that both nouns in a noun-noun-verb triple are amenable to a subject interpretation, but that one of them is a more likely subject than the other. Then, it is reasonable to expect the model to process the less likely subject candidate as the object of the verb in the triple. Another promising line of development is based on the observation that the order in which verb arguments appear in context is also lexically governed: in Italian, for example, report verbs show a strong tendency to select subjects post-verbally. Dell'Orletta *et al.* (2005) report a substantial improvement on the model performance on Italian *SOI* when lexical information is taken into account, as a lexicalized MaxEnt model appears to integrate general constructional and semantic biases with lexically-specific preferences. In a cross-lingual perspective, comparable evidence of lexical constraints on word order would allow us to discover language-wide invariants in the lexicon-grammar interplay.

References

Bates E., MacWhinney B., Caselli C., Devescovi A., Natale F., Venza V. 1984. A crosslinguistic study of the development of sentence interpretation strategies. *Child Development,* 55: 341-354.

Bohmova A., Hajic J., Hajicova E., Hladka B. 2003. The Prague Dependency Treebank: Three-Level Annotation Scenario, in A. Abeille (ed.) *Treebanks: Building and Using Syntactically Annotated Corpora,* Kluwer Academic Publishers, pp. 103-128.

Bybee J. 2002. Sequentiality as the basis of constituent structure. in T. Givón and B. Malle (eds.) *The Evolution of Language out of Pre-Language*, Amsterdam: John Benjamins. 107-132.

Croft W. 2003. *Typology and Universals. Second Edition*, Cambridge University Press, Cambridge.

Bresnan J., Dingare D., Manning C. D. 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. *Proceedings of the LFG01 Conference*, Hong Kong: 13-32.

Dell'Orletta F., Lenci A., Montemagni S., Pirrelli V. 2005. Climbing the path to grammar: a maximum entropy model of subject/object learning. *Proceedings of the ACL-2005 Workshop "Psychocomputational Models of Human Language Acquisition"*, University of Michigan, Ann Arbour (USA), 29-30 June 2005.

Hay J., Baayen R.H. 2005. Shifting paradigms: gradient structure in morphology, *Trends in Cognitive Sciences*, 9(7): 342-348.

Hayes B. 2000. Gradient Well-Formedness in Optimality Theory, in Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer (eds.) *Optimality Theory: Phonology, Syntax, and Acquisition*, Oxford University Press, pp. 88-120.

Lenci A. *et al.* 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13 (4): 249-263.

MacWhinney B. 2004. A unified model of language acquisition. In J. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, Oxford University Press, Oxford.

Manning C. D. 2003. Probabilistic syntax. In R. Bod, J. Hay, S. Jannedy (eds), *Probabilistic Linguistics*, MIT Press, Cambridge MA: 289-341.

Miyao Y., Tsujii J. 2002. Maximum entropy estimation for feature forests. *Proc. HLT2002*.

Montemagni S. *et al.* 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) *Treebanks. Building and Using Parsed Corpora*, Kluwer, Dordrecht: 189-210.

Ratnaparkhi A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. Dissertation, University of Pennsylvania.

Constraints for S		
Feature	Italian	Czech
Preverbal	1.31E+00	1.24E+00
Nomin	na	1.23E+00
Agr	1.28E+00	1.18E+00
Anim	1.28E+00	1.16E+00
Inanim	8.16E-01	1.03E+00
Postverbal	5.39E-01	8.77E-01
Genitive	na	2.94E-01
NoAgr	1.52E-01	7.71E-02
Dative	na	2.85E-02
Accus	na	8.06E-03
Instrum	na	3.80E-03

Constraints for O		
Feature	Italian	Czech
Genitive	na	1.51E+00
NoAgr	1.58E+00	1.50E+00
Dative	na	1.49E+00
Accus	na	1.39E+00
Instrum	na	1.39E+00
Postverbal	1.38E+00	1.17E+00
Inanim	1.23E+00	9.63E-01
Agr	4.67E-01	6.67E-01
Anim	3.17E-01	6.63E-01
Preverbal	2.11E-02	5.40E-01
Nomin	na	2.22E-02

Table 5 – *Ranked constraints for S and O in Czech and Italian*